

HP Extended Cluster for RAC—100 kilometer separation becomes a reality



Executive summary.....	2
Business problems.....	2
The solution.....	3
Context for the solution	3
Solution overview.....	3
Solution components.....	4
100 km Extended Cluster for RAC—what’s new	6
The tests	7
Testing methodology	7
Test architecture	8
Supported hardware configuration	8
Supported software configuration	8
Test descriptions.....	8
IPC test description	9
I/O test description	9
OLTP test description.....	9
Failover test description.....	9
Test results and tuning.....	9
The next steps.....	10
Solution flexibility	10
Conclusions	11
References	12
For more information.....	12

Executive summary

The management and mitigation of risk within an IT environment has always been one of the primary responsibilities of IT management. This task has traditionally been addressed by the duplication of critical computing components, the elimination of single points of failure, and the potential use of a secondary facility to continue operations in the event of a catastrophic loss of a primary data center.

Although this approach visibly demonstrates diligence toward risk management, it is achieved by deploying a set of duplicated and underutilized resources. HP Extended Cluster for RAC combines the high-availability product HP Serviceguard Extension for RAC and the disaster-tolerant solution HP Extended Cluster to specifically address the reduction of risk and to capitalize on the IT investment in any environment that uses an Oracle9i Real Application Clusters (RAC) data repository.

HP Extended Cluster for RAC allows for a single database instance to be split across two data centers, separated by an unprecedented 100 kilometers. Because the two sites are functioning as a virtual single entity, all resources can be utilized at all times. The ability to take distributed IT components and meld them into a homogeneous resource pool is made possible by the HP Virtual Server Environment (VSE) portfolio of solutions.

HP is renowned for providing world-class high-availability and disaster-tolerant solutions. With the addition of virtualization, it is now possible to fully utilize all IT resources throughout an enterprise and still maintain previous levels of availability and tolerance. During normal operations, users enjoy the power of harnessing the organization's complete suite of computing resources. However, in the event of the loss of an entire data center, the remaining data center would continue to function. The 100 km separation dramatically diminishes any likelihood of a single disaster impacting more than one of the data centers at any one time.

In addition to the replication of data between two data centers, data is also being synchronized, and the entire computing environment is able to function as a single virtual entity. Even though there may be up to 100 kilometers between the data centers, the administration of the application and data is equivalent to managing a traditional RAC application located in a single data center.

With the Virtual Server Environment, HP has created a unique solution that provides unrivaled levels of risk management and return on IT investment, and it has simultaneously increased each enterprise's ability to rapidly accommodate and capitalize on volatile business conditions.

HP has made extending a RAC solution a reality.

Business problems

If a group of CIOs was asked to identify their primary cause of stress and worry, the majority would answer "controlling IT risks and costs." Indeed, many would say that these are the top two most important challenges within a CIO's charter.

In striving to balance risk management and fiscal responsibilities, every IT manager is faced with a powerful dichotomy. Risk can be reduced by investing in duplication and redundancy, and costs can be driven down by exposing the organization to increased levels of risk. Therefore, simultaneously lowering both risk and cost has traditionally been a highly elusive quest.

A previously acceptable and widely adopted approach to risk management was for an IT organization to purchase excess equipment to create a pool of surplus CPU, storage, and connectivity resources. This method works for dealing with many adverse situations. However, in today's harsh business climate, with the microscopic examination of every dollar spent, it represents a highly visible example of poor utilization of IT assets.

To allow for changes in business conditions, IT system configurations are typically sized to handle maximum anticipated loads and designed to be scalable, with additional hardware. This approach provides a fair degree of flexibility, but it does so at the expense of asset utilization. HP studies have shown that, on average, system use ranges from 30 to 50 percent of available resources, and yet it is not unusual to see a number of applications bottlenecked by resource constraints.

HP has been providing solutions to help maximize clients' return on IT investments for many years. By combining newly created technologies with proven industry-standard solutions, HP can now demonstrate the next generation of offerings that increase utilization, manage risk, and make the most of IT expenditures.

The solution

Context for the solution

The Adaptive Enterprise is HP's vision for helping customers synchronize business and IT to capitalize on change. Virtualization lets you balance two seemingly contradictory areas—cost and agility—by pooling and sharing servers, storage, networking, and other infrastructure devices and allocating them across applications and processes as your business demands them. The new adaptive enterprise has management capabilities to sense change in business demand and trigger the dynamic supply of virtualized resources.

HP defines virtualization as managing an IT environment as a single entity by pooling and sharing resources so that supply automatically meets demand in real time. This holistic approach allows the entire IT resource pool to be viewed as a single virtual entity that can be directed—by business priorities and policies—to dynamically serve the requirements of the enterprise.

In isolation, virtualization over extended distances has the potential to actually increase the burden placed on IT organizations. However, by combining virtualization with powerful real-time automation tools, HP has been able to create a compelling suite of solutions that release previously underutilized assets without causing additional IT overhead.

HP has included the nucleus of its intelligent virtualization concept in the HP Virtual Server Environment (VSE) portfolio of solutions. Each of the VSE offerings is fundamentally designed to achieve an identical set of objectives:

- Improve return on IT investment (RoIT)
- Provide increased abilities to accommodate and exploit business volatility
- Enable elevated levels of IT service to be provided to the enterprise
- Present unrivaled choices for the management and mitigation of risk

Solution overview

At the very center of the "risk versus cost" challenge is the question "How can I increase application availability and resiliency while also aggressively managing costs?" The concept of using multiple data centers, with replicated system configurations, to provide fallback capabilities through redundancy is well proven, but not new. This approach minimizes the impact of the total loss of one data center, but it achieves this through the gross underutilization of IT assets.

HP has addressed this issue with a unique and compelling blend of highly utilized, high-availability (HA) and disaster-tolerant solutions (DTS). HP has retained all the advantages of split data centers and is demonstrating solutions that virtually eliminate the presence of redundant, unnecessarily duplicated or underdeployed resources. This includes support for Oracle® Real Application Clusters (RAC), designed to supplement the Oracle9i data repository, which is one of the world's most popular databases.

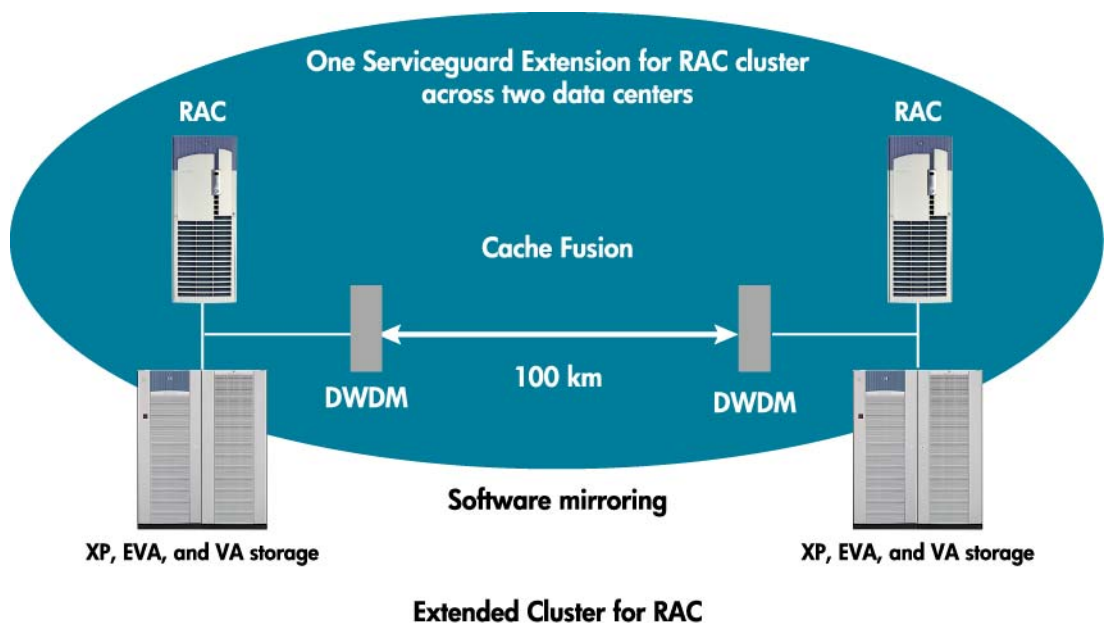
Oracle9i RAC allows multiple instances to access a single logical database, across multiple servers, with all nodes able to concurrently execute transactions against the same data repository. However, in order for the systems to be positioned in separate data centers, several other key solution components have to be added. These include data center connectivity, workload balancing, and a variety of service managers—for cluster arbitration, business policy enforcement, volume management, partitioning, mirroring and synchronization, etc.—as well as supporting hardware.

To publicly demonstrate the real-world feasibility of hosting a single logical database instance across multiple discrete data centers separated by multiple kilometers, HP enhanced the HP Serviceguard Extension for RAC (SGeRAC) solution.

Solution components

The highly pervasive RAC configuration traditionally offers strong high-availability and scalability characteristics. SGeRAC delivers the ability to perform identical operations across two remotely located data centers and gives the configuration increased stability through additional high availability and disaster tolerance. In addition, it provides efficiency, with the excellent utilization of all resources, which can be dynamically allocated.

Figure 1. Extended Cluster for RAC—100 km

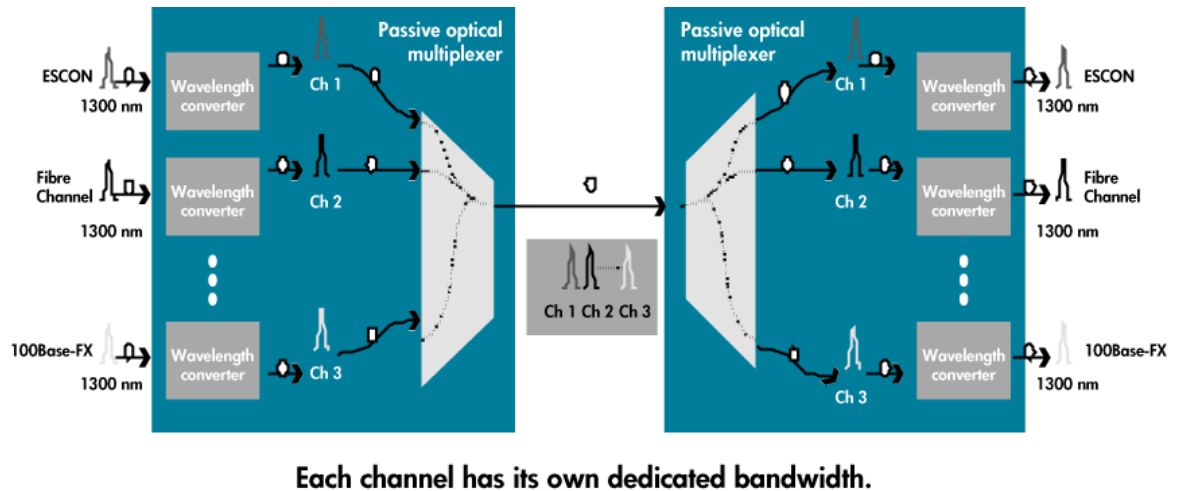


Cache Fusion is a key component of RAC that uses cluster interconnect technologies to facilitate virtual buffer sharing through the use of direct memory accessing. The Cache Fusion architecture creates a single virtual cache across all the nodes of the cluster. The resulting single view of all buffer caches minimizes disk I/O by allowing any database request to be served by any node in the cluster.

The introduction of a second data center creates many advantages for risk management and mitigation. However, until virtualization with SGeRAC was offered, it also had the potential to introduce unacceptable levels of asset underutilization, operational complexity, and transaction latency.

The Extended Cluster for RAC architecture relies on the same components as the co-located system configuration, but it introduces several additional key technologies. System-to-system communication is now facilitated by the addition of a pair of Dense Wavelength Division Multiplexing (DWDM) devices connected via dark fiber. Database replication is achieved via host-based software mirroring, with RAC synchronizing the database caches via Cache Fusion.

Figure 2. DWDM: Dense Wavelength Division Multiplexing



DWDM employs optoelectronic technology to simultaneously transmit multiple separate optical signals through a single optical fiber. This is accomplished by changing the wavelength of the incoming optical signals to permit co-existence on a common fiber¹.

The transmitting DWDM device is able to multiplex multiple converted optical inputs over the same fiber optic cable. The destination DWDM reverses the process; it de-multiplexes the signals and converts them back to their original wavelength. This allows network, disk, and application communications to share the same DWDM link and also provides the capability to extend the Fibre Channel link distances. Examples shown in Figure 2 include the multiplexing and de-multiplexing of Fibre Channel, 100Base-FX, and ESCON (Enterprise System Connection—a large-system I/O channel architecture) transmissions.

The DWDMs and fiber optic media together provide the very low latency interconnect that is essential to RAC performance and scalability.

Dark fiber is existing fiber optic cable that has not yet been lit. This excess fiber capacity is very frequently already available, and using it to connect DWDM devices is rapidly gaining popularity as a cost-effective mechanism to achieve increased bandwidth. Because the conduit is typically dedicated to this solution, it is permanently able to offer full capacity and security to the associated transmissions.

¹ An in-depth description of the evolution, testing, and deployment of DWDM can be found in the document titled *DWDM: A White Paper*, by Joseph Algieri and Xavier Dahan.

The complex tasks of managing and linking business priorities to the appropriate resource allocations are handled by the highly integrated components of the HP Virtual Server Environment portfolio of solutions:

- HP Serviceguard and Serviceguard Extension for RAC (SGeRAC)
- HP-UX Workload Manager (WLM)
- HP Partitioning Continuum
- HP Instant Capacity on Demand (iCOD)

HP SGeRAC brings data protection, application availability, and ease of management to servers and server partitions to create an enterprise cluster that delivers highly available application services to LAN-attached clients. HP Serviceguard is able to monitor the health of each node and respond to failures in a way that minimizes or eliminates application downtime. When used in conjunction with SGeRAC, it allows for rapid and transparent recovery from LAN and application failures, while still maintaining scalability, data integrity, and configuration flexibility.

HP-UX Workload Manager provides automatic resource allocation and application performance management through the use of prioritized service-level objectives (SLOs). It is the key to enabling applications to be stacked. Goal-based resource management provides highly predictable response times for all mission-critical applications, regardless of their location. The tight integration of HP Serviceguard and WLM allows CPUs to be assigned to specific HP Serviceguard packages after a failover.

HP Partitioning Continuum components are the industry's broadest range of hard and virtual partitions, and they offer partitioning tools that help provide resource flexibility, improved system utilization, and lowered costs in consolidated environments.

HP Instant Capacity on Demand is an innovative mechanism to rapidly respond to unpredictable business volatility through the deployment of preconfigured systems that provide additional computing capacity—but clients are only charged when the additional resources are utilized.

100 km Extended Cluster for RAC—what's new

The solution components work together to provide the best aspects of the HP Virtual Server Environment, HP enterprise clusters, and Oracle Real Application Clusters: comprehensive resource utilization, high availability, data integrity, scalability, and reduced administration costs. Compelling on their own, these characteristics are further complemented by the ability to split databases across multiple data centers, providing full disaster tolerance and unprecedented levels of risk mitigation and cost effectiveness.

Utilizing its intelligent policy engine, the Workload Manager (WLM) component of the VSE allows the real-time allocation of assets throughout the virtual resource pool—yielding enhanced server utilization, driving increased return on IT investments, and dramatically improving the enterprise's ability to accommodate business volatility.

If an entire data center is lost, the remaining data center continues to function through the re-routing of users to the functioning environment. This capability provides continuous availability across the two data centers. Once a failed data center comes back online, all resynchronization takes place automatically and is transparent to users.

The administration of the overall environment is greatly simplified because the application is resident on a single data repository. The concept of unnecessarily replicated databases, with the associated burden of replicated management chores, has all but been eliminated. Even when spread across a distance of 100 kilometers, the Oracle9i database is still a single database instance, possessing inherent economies of system administration over multi-instance solutions.

The move to 100 km separation represents a dramatic increase in overall application resiliency. The ability to place such a large distance between two data centers assures that only the most widespread disaster will impact more than one of the installations.

The tests

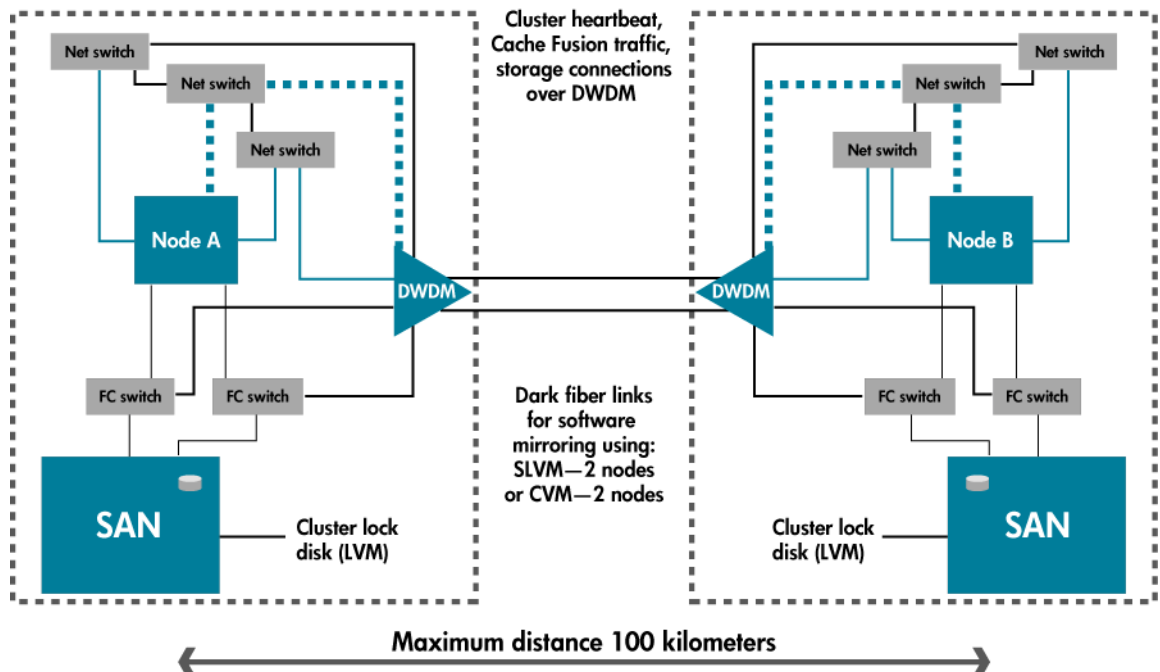
Testing methodology

HP's years of experience with Extended Cluster and SGeRAC were leveraged in developing the test plans for the extended distance testing. Together with partners Oracle, AT&T, and Nortel, HP's testing focused on demonstrating the ability to achieve a robust solution capable of maximizing resource utilization across extended distances with specific attention paid to high availability, disaster tolerance, and performance characteristics between remote data centers separated by distances up to 100 km.

The underlying premise for the testing was the validation of the remotely distributed cluster's ability to sustain full functionality while being subjected to failovers. Testing was done at distances of 25, 50, and 100 km to validate availability, performance, and network latency.

Once the data center to data center link was established, a series of loading and failure scenarios was executed to validate the expected disaster-tolerant characteristics of the configuration. The scenarios simulated failure-inducing events on a loaded configuration that traditionally would prove catastrophic in a non-SGeRAC environment. Each test was structured to induce significant transaction-based traffic, performing the sequenced set of failure-inducing events and the subsequent validation of data integrity during and following the failure.

Figure 3. Two data center Extended Clusters for RAC—up to 100 km currently supported configuration



Test architecture

The test configuration for the 100 km evaluation followed the supported configuration for all subsequent SGeRAC implementations.

Supported hardware configuration

- 2-node cluster (PA-RISC or HP Integrity servers, not mixed)
 - 2- or 4-node cluster is supported up to 10 km with VERITAS Cluster Volume Manager (CVM)
- Redundant heartbeat plus 1 or 2 LANs for Cache Fusion traffic
 - Clusters with VERITAS CVM; only single heartbeat subnet is supported (use primary/standby LANs)
- Any Fibre Channel storage (using VA, XP, EVA, or JBOD technologies)
 - Note: EVA disk arrays do not currently support VxVM/CVM
- Both TCP/IP networking and Fibre Channel data can go through the same DWDM box; redundant DWDM boxes in each data center are not required if the box is designated as acceptably fault tolerant
- At least two dark fiber links between data centers
 - During testing, AT&T-supplied dark fiber optic links were used between data centers (to further elevate disaster-tolerance characteristics, alternate physical routes can be added to remove single points of failure)
 - Note: TCP/IP networking and Fibre Channel data pass through the same DWDM box; in a non-test environment, redundant DWDM boxes in each data center are not required if fault tolerant design specifications meet disaster-tolerance requirements
- Cluster arbitration: dual cluster lock disks (LVM volume groups) or Quorum Server
 - Note: Three site configuration is supported using Quorum Server to prevent split-brain
- Nortel OPTera DWDM switches were used for initial testing and certification

Supported software configuration

- Operating system: HP-UX 11.11 or HP-UX 11i v2.0
- Shared Logical Volume Manager (with physical volume links and HP Mirrordisk/UX, HP StorageWorks SecurePath with EVA storage)
- VERITAS VxVM/CVM 3.5 (with DMP and mirroring) for cluster volume management
- HP Serviceguard Quorum Server or dual cluster lock disks for cluster arbitration (the dual cluster lock disks are required—one at each center—to facilitate recovery from an entire data center failure)
- HP Serviceguard and HP Serviceguard Extension for RAC A.11.15 or later
- HP Serviceguard OPS edition A.11.14
- Oracle RAC 9.2
- HP StorageWorks Extended Fabric on Fibre Channel switches (StorageWorks Extended Fabric enables dynamically allocated long distance configurations in a Fibre Channel switch)

Test descriptions

The following categories of testing were performed:

- IPC tests
- I/O tests
- Online transaction processing (OLTP)-like workload tests
- Failover tests

IPC test description

Raw IPC throughput was evaluated using CRTEST, a micro-level performance benchmark. The test first updates a set of blocks in a hash cluster table on instance A, and then an increasing number of clients running SELECT are started on instance B. These queries cause messages to be sent from instance B to instance A; instance A returns a CR block. CR Fairness Down Converts were disabled for this test to create a fundamental dialog of “send a message” and “receive a CR block” back.

The CRTEST (IPC) tests were performed initially using one Gigabit Ethernet network for cluster interconnection, and then the tests were repeated using two interconnects.

I/O test description

The I/O tests were executed using the Diskbench (db) disk subsystem performance measurement tool. The db tool measures the performance of a disk subsystem, host bus adapter, and driver in terms of throughput (for sequential operation) and number of I/Os (for random operation). Diskbench can evaluate the performance of kernel drivers and can be used on one-way or multiprocessor systems to completely saturate the processors and effectively measure the efficiency of a disk subsystem and associated drivers. The tests were performed using a mix of 60 percent reads and 40 percent writes to emulate “real-world” traffic.

OLTP test description

The industry-standard TPC-C test was utilized to simulate OLTP transactions over the cluster interconnects and DWDM network. TPC-C is able to emulate multiple transaction types and complex database structures. The benchmark involves a mix of five concurrent transactions of varying type and complexity that were executed online and queued for deferred execution. TPC-C is widely acknowledged as providing one of the most realistic loading simulations within a comprehensive computing environment.

Failover test description

Multiple tests were performed to simulate the wide variety of scenarios that could impact a data center. Host failure, storage device failure, DWDM link failure, and catastrophic data center failure were all emulated to stress the configuration in the most rigorous and realistic manner possible. Intra data center connections to storage, either via the Interswitch Linking (ISL) switch or directly to an array, were removed one at a time to simulate localized failures. Characteristics of the configuration’s response to each event were observed, including the reaction of the Oracle instance, and behavior during resynchronization of data volumes as connections became re-established was also watched. Data integrity and user impact were closely monitored throughout all phases of testing.

Test results and tuning

The most critical result found in the testing is that the 100 km Serviceguard Extension for RAC solution, running over DWDM, fully demonstrates the complete set of high-availability characteristics seen on a co-located HP Serviceguard cluster, allowing the server and application cluster to continue to perform and be accessible when failure occurs on one or more of the components. Data integrity was maintained for the duration of all testing.

For host failure emulation, a server was shut down, making the Oracle database components resident on that system inaccessible for two to three minutes. All traffic was able to move to the second data center. Once restarted, the failed system resumed operations.

A forced single storage device failure had no impact on the Oracle instance. Even though the test was continued for several minutes, no visible user impact or loss of data integrity was observed.

DWDM link failure did not compromise the cluster, which continued to stay up for the duration of the test cycle. No noticeable negative behaviors in HP Serviceguard and SGeRAC performance or functionality were noted.

The solution is able to perform strongly against stringent proprietary and industry-standard benchmark tests, still providing the enterprise with unprecedented levels of disaster tolerance and resource utilization across the 100 km link. Even during recovery, the full functionality of the Oracle repository remained available.

As expected, in the un-optimized configuration, IPC, I/O, and application operations did show a degree of performance degradation related to separation distances. This indicates that assessment of the application and workload characteristics of the target environment is critical to making the most of the implementation over extended distances. During the evaluation, informal manipulation clearly demonstrated that tuning and hardware selection can have a significant impact on overall system performance.

It was found that the assignment of buffer credits within the DWDM devices had a significantly positive impact on throughput performance of the application, as shown in this table:

	27 Buffer credits		60 Buffer credits	
	Read	Write	Read	Write
Local	94 mb/s	90 mb/s	94 mb/s	90 mb/s
50 km	50 mb/s	25 mb/s	94 mb/s	55 mb/s
100 km	25 mb/s	12 mb/s	50 mb/s	27 mb/s

The use of multiple DWDM channels allows each to have its own unimpeded bandwidth, subject to the total available aggregated bandwidth. Further increases in performance were observed with the addition of extra interconnects to distribute Cache Fusion traffic and I/O cards to distribute I/O loads. The utilization of switches with more buffer credits allows full bandwidth to be maintained at longer distances.

The next steps

A variety of performance enhancements at various levels within the Extended Cluster for RAC architectural solution are available, such as network buffering, workload distribution, and database partitioning. In many cases, these enhancements will be specific to individual environments. As these solutions are deployed, a collection of best practices will be shared.

In order to continue providing increasing business value, HP is constantly investing in creating new solutions and further enhancing existing portfolios. A demonstration of this commitment will be in the capability to handle numbers of nodes even higher than the current specifications.

During the evaluation and testing, informal manipulation clearly demonstrated that tuning and hardware selection can have a significant impact on overall system performance. In addition, the ability to extend beyond 100 km is being explored with key networking partners.

Solution flexibility

The 100 km Extended Cluster for RAC solution was tested using the components detailed above. However, the configuration is extremely accommodating in the choice of acceptable hardware and software. This ability allows customers to leverage existing purchases and to rapidly move to an operational implementation without the unnecessary purchase of a completely new infrastructure.

During the testing period, a wide range of storage devices was introduced into the configuration. The majority of HP-UX enterprise server-compatible storage devices will perform suitably in the Extended Cluster for RAC environment, from the high-end HP StorageWorks XP models to the low-cost HP EVA storage solution based on Dynamic Provisioning Architecture (DPA).

In addition, HP Serviceguard Quorum Server can be introduced to provide cluster lock management for tie-breaking and autonomous clustering following any failure that impacts cluster integrity.

There are a variety of options for load balancing via access clients, including Resonate's Central Dispatch product and a selection of hardware-based choices from Cisco.

Given the tremendous transparency of the DWDM devices, it is possible to run a wide variety of protocols across the fiber. During testing, Gigabit Ethernet was used for the system-to-system connectivity. Network switches can be 100Base-T (TX or FX), 1000Base-T (TX or FX), or FDDI. The connections between the network switches and the DWDM boxes must currently remain fiber optic.

As demonstrated by the Oracle9i RAC implementation, HP is introducing Virtual Server Environments enhanced for specific application server and database environments. HP released the Virtual Server Environment for BEA WebLogic Server on HP-UX, which provides the most effective way for customers to quickly implement virtualization within their application server and database environments to achieve the benefits of an adaptive enterprise.

Conclusions

HP has successfully shown the viability of Extended Cluster for RAC using 100 km data center separations. It again leads the industry by demonstrating the power of an HP Serviceguard Extension for RAC configuration split across two discrete data centers separated by 100 kilometers. The 100 km distance significantly increases the probability of any single disaster only impacting one data center.

Traditionally, disaster tolerance has been accomplished by an environment dedicated to "active standby"—resources are not utilized until needed. The 100 km RAC solution destroys this paradigm—there is now full utilization of all resources in all data centers at all times. Data is being replicated between the two data centers and is fully synchronized, allowing location-independent access to the applications. The Virtual Server Environment is permitting both data centers to ostensibly see the data simultaneously and to continually harness the aggregated resources from both locations.

HP is able to deliver a highly reliable and proven solution based on HP Serviceguard, VSE, SGeRAC, and Oracle9i RAC solutions. This combination is already widely purchased and accredited in the industry today. The demonstrated ability to now further mitigate risk by physically separating data centers by distances of up to 100 km brings unprecedented levels of return on IT investment and risk management.

HP has made extending a RAC solution a reality.

References

DWDM: A White Paper, by Joseph Algeri and Xavier Dahan, available on request from HP.

For more information

www.hp.com/go/ha

www.hp.com/go/virtualization

© 2004 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Oracle is a registered U.S. trademark of Oracle Corporation, Redwood City, California.

5982-3575EN Rev. 1, 03/04/2004

